In the claims:

1-164. (Canceled)

165. (New) A method of extracting significant patterns from a dataset having a plurality of sequences defined over a lexicon of tokens, the method comprising, for each sequence of the plurality of sequences: searching for partial overlaps between said sequence and other sequences of the dataset, applying a significance test on said partial overlaps, and defining a most significant partial overlap as a significant pattern of said sequence, thereby extracting significant patterns from the dataset.

166. (New) The method of claim 165, wherein said search for partial overlaps is by constructing a graph having a plurality of paths representing the dataset and searching for partial overlaps between paths of said graph.

167. (New) The method of claim 166, wherein said search for partial overlaps between paths of said graph comprises:

defining, for each path, a set of sub-paths of variable lengths, thereby defining a plurality of sets of sub-paths; and

for each set of sub-paths, comparing each sub-path of said set with sub-paths of other sets.

168. (New) The method of claim 166, wherein said graph comprises a plurality of vertices, each representing one token of the lexicon, and further wherein each path of said plurality of paths comprises a sequence of vertices respectively corresponding to one sequence of the dataset.

169. (New) The method of claim 166, further comprising calculating, for each path, a set of probability functions characterizing said partial overlaps.

170. (New) The method of claim 165, further comprising grouping at least a few tokens of said significant pattern, thereby redefining the dataset.

171. (New) The method of claim 165, wherein the dataset comprises a corpus of text.

172. (New) The method of claim 165, wherein the dataset comprises a protein database.

173. (New) The method of claim 165, wherein the dataset comprises a DNA database.

174. (New) The method of claim 165, wherein the dataset comprises an RNA database.

175. (New) The method of claim 165, wherein the dataset comprises a recorded speech.

176. (New) The method of claim 165, wherein the dataset comprises a corpus of music notes.

177. (New) The method of claim 165, wherein the dataset comprises a weblog database.

178. (New) The method of claim 165, wherein the dataset comprises trajectory records of a transportation network.

179. (New) The method of claim 165, wherein the dataset comprises activity records of a self-active system.

180. (New) The method of claim 165, wherein the dataset comprises records of operational steps in a technical process.

181. (New) A method of generalizing a dataset having a plurality of sequences defined over a lexicon of tokens, the method comprising:

searching over the dataset for similarity sets, each similarity set comprising a plurality of segments of size L having L-S common tokens and S

uncommon tokens, each of said plurality of segments being a portion of a different sequence of the dataset; and

defining a plurality of equivalence classes corresponding to uncommon tokens of at least one similarity set, thereby generalizing the dataset.

182. (New) The method of claim 181, wherein said definition of said plurality of equivalence classes comprises, for each segment of each similarity set:

extracting a significant pattern corresponding to a most significant partial overlap between said segment and other segments or combination of segments of said similarity set, thereby providing, for each similarity set, a plurality of significant patterns; and

using said plurality of significant patterns for classifying tokens of said similarity set into at least one equivalence class;

thereby defining said plurality of equivalence classes.

183. (New) The method of claim 182, further comprising, prior to said search for said similarity sets:

extracting a plurality of significant patterns from the dataset, each significant pattern of said plurality of significant patterns corresponding to a most significant partial overlap between one sequence of the dataset and other sequences of the dataset; and

for each significant pattern of said plurality of significant patterns, grouping at least a few tokens of said significant pattern, thereby redefining the dataset.

184. (New) The method of claim 181, further comprising, for each similarity set having at least one equivalence class, grouping at least a few tokens of said similarity set thereby redefining the dataset.

185. (New) The method of claim 181, further comprising for each sequence, searching over said sequence for tokens being identified as members of previously defined equivalence classes, and attributing a respective equivalence class to each identified token, thereby generalizing said sequence, thereby further generalizing the dataset.

186. (New) The method of claim 183, further comprising constructing a graph having a plurality of paths representing the dataset, wherein each extraction of significant pattern is by searching for partial overlaps between paths of said graph.

187. (New) An apparatus for generalizing a dataset having a plurality of sequences defined over a lexicon of tokens, the apparatus comprising:

(a) a searcher, for searching over the dataset for similarity sets, each similarity set comprising a plurality of segments of size L having L-S common tokens and S uncommon tokens, each of said plurality of segments being a portion of a different sequence of the dataset; and

(b) a definition unit, for defining a plurality of equivalence classes corresponding to uncommon tokens of at least one similarity set, thereby generalizing the dataset.

188. (New) The apparatus of claim 187, further comprising an extractor, capable of extracting, for a given set of sequences, a significant pattern corresponding to a most significant partial overlap between one sequence of said set of sequences and other sequences of said set of sequences, thereby providing, for said given set of sequences, a plurality of significant patterns.

189. (New) The apparatus of claim 188, wherein said given set of sequences is a similarity set, hence said plurality of significant patterns corresponds to said similarity set.

190. (New) The apparatus of claim 188, wherein said classifier is designed for selecting a leading significant pattern of said similarity set, and defining uncommon tokens of segments corresponding to said leading significant pattern as an equivalence class.

191. (New) The apparatus of claim 188, wherein said given set of sequences is the dataset, hence said plurality of significant patterns corresponds to the dataset.

192. (New) The apparatus of claim 188, further comprising a first grouper for grouping at least a few tokens of each significant pattern of said plurality of significant patterns.

193. (New) The apparatus of claim 187, further comprising a second definition unit having a second searcher, for searching over each sequence for tokens being identified as members of previously defined equivalence classes, wherein said second definition unit is designed to attribute a respective equivalence class to each identified token.

194. (New) The apparatus of claim 188, further comprising a constructor, for constructing a graph having a plurality of paths representing the dataset.